Reproducibility of 18F-FDG and 3'-deoxy-3'-18F-fluorothymidine PET tumor volume measurements.

Mathieu Hatt, Catherine Cheze-Le Rest, Eric O. Aboagye, Laura M. Kenny, Lula Rosso, Federico E. Turkheimer, Nidal M. Albarghach, Jean-Philippe Metges, Olivier Pradier, Dimitris Visvikis

Reproducibility of 18F-FDG and 18F-FLT PET tumor volume measurements

Mathieu Hatt[1], Catherine Cheze-Le Rest[1,2], Eric O. Aboagye[3], Laura M. Kenny[3], Lula Rosso[3], Federico E. Turkheimer[3], Nidal M. Albarghach[1,4], Jean-Philippe Metges[4], Olivier Pradier[1,4], Dimitris Visvikis[1]

1. INSERM, U650, LaTIM, CHU Morvan, Brest, France

2. Academic Department of Nuclear Medicine, CHU Morvan, Brest, France

3. MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital, London, UK

4. Institute of Oncology, CHU Morvan, Brest, France

Short running title: PET volume determination reproducibility

Corresponding author:

Mathieu HATT
LaTIM, INSERM U650
CHU MORVAN
5 avenue Foch
29609 Brest
France

e-mail: hatt@univ-brest.fr

Tel.:+33298018111

Fax: +33298018124

Wordcount: 5999

Abstract

The objective of this study was to establish the repeatability and reproducibility limits of several volume related PET image derived indices, namely tumour volume (TV), SUV_{mean}, and Total Glycolytic or Proliferative Volume (TGV, TPV), relative to that of SUV_{max}, commonly employed in clinical practice

Methods: Fixed and adaptive thresholding, fuzzy C-means (FCM) and fuzzy locally adaptive bayesian (FLAB) were considered for TV delineation. Double baseline 18-FDG (17 lesions, 14 esophageal cancer patients) and 18-FLT (12 lesions, 9 breast cancer patients) PET scans acquired at a mean of 4 days interval and prior to any treatment were used for reproducibility evaluation. The repeatability of each method was evaluated on the same datasets and compared to manual delineation.

Results: A negligible variability of <5% was measured for all segmentation approaches in comparison to manual delineation (5%-35%). SUV_{max} reproducibility levels were similar to others previously reported with a mean percentage difference of 1.8%±16.7% and -0.9%±14.9% for the FDG and FLT lesions respectively. The best TV and TGV/TPV reproducibility limits ranged from -21% to 31% and -30% to 37% for FDG and FLT images respectively, whereas the worst reproducibility limits ranged from -90% to 73% and -68% to 52% respectively.

Conclusion: The reproducibility of estimating TV, SUV_{mean} and derived TGV/TPV was found to vary among segmentation algorithms. Some differences between FDG and FLT scans were observed mainly due to differences in overall image quality. The smaller reproducibility limits for volume derived image indices were similar to those for SUV_{max}, suggesting that the use of appropriate delineation tools should allow determining tumor functional volumes in PET images in a repeatable and reproducible fashion.

1. Introduction

Most of current PET clinical practice for diagnosis, staging, prognosis, therapy response assessment and patient follow-up rely on manual and visual analysis *(1)*. The index most commonly employed in PET clinical studies is the standardized uptake value (SUV). In order to obtain this index of activity accumulation a region of interest (ROI) should be determined, usually drawn manually or using some fixed threshold. Despite not being the only factor that can affect the accuracy of SUVs, the type and size of ROI is a large contributor to the variability of such measures as has been previously demonstrated *(2,3)*. A popular alternative is the use of the pixel with the maximum activity value, usually referred to as SUV_{max}. A large number of studies have demonstrated its prognostic and predictive value, despite the fact that it is sensitive to image noise *(4,5)*. On the other hand, there are a limited number of mostly recent studies that have explored the use of overall tumor volume (TV) as an index for prognosis and response assessment *(6-8)* considering either the TV alone or in combination with the mean SUV (SUV_{mean}) to form the total glycolytic or proliferative (for FDG and FLT respectively) volume (TGV/TPV), defined as the product of TV x SUV_{mean} *(9-11)*.

Clearly one of the issues associated with the use of functional volumes derived from PET images, which is directly responsible for the reduced use of such indices, is the accuracy, robustness, repeatability and reproducibility considering the delineation. On the one hand, manual delineation of functional volumes using PET images leads to high inter- and intra-observer variability *(3)*, principally arising from the poor quality of PET images. On the other hand, current state-of-the-art algorithms for functional volume segmentation consist of fixed *(12)* or adaptive thresholding approaches *(13,14)*. Although attractive as a result of the easiness of use, the

drawbacks of fixed threshold approaches are numerous, since the value of the threshold to be used for each lesion clearly depends on multiple factors, such as lesion contrast and size as well as image noise *(15)*. The solutions based on the use of adaptive thresholding consider the contrast between the object to delineate and its surrounding background. However, they require imaging system specific optimisation carried out considering uniformly filled spherical lesions reducing the robustness of the approach, particularly in the case of multi-centre trials. In addition, their performance depends on the background ROI choice which can in turn lead to reduced inter-observer reproducibility for functional volume determination. A few automatic algorithms have been recently proposed *(16-19)*. The main difference between these algorithms and the threshold-based approaches is that they automatically estimate the parameters of interest and find the optimal regions' characteristics in a given image without system-dependent parameters. This may allow reducing issues associated with deterministic approaches based on thresholding, potentially increasing the robustness and reproducibility of PET functional volumes determination *(20)*.

Establishing the level of reproducibility and repeatability is essential in the use of any image derived index and its use in prognostic or therapy response studies, allowing the evaluation of which change between two studies can be considered significant. To date there have only been a limited number of reproducibility studies *(21-25)*, almost exclusively concentrating on SUV_{max} and SUV_{mean} variability in double baseline FDG PET scans, showing a relative absolute percentage difference of up to 13% with a standard deviation of 10%. The reproducibility of quantitative indices (Patlak influx constant, Ki), associated with the acquisition of dynamic datasets, have been also assessed *(21,22)* showing similar levels of reproducibility

(mean percentage difference of 8%-10%). Studies on the reproducibility of such indices in the case of FLT PET imaging has shown that changes larger than 15-20% and 25-30% may be considered significant in SUV_{mean} (obtained using a 41% fixed threshold) and SUV_{max} or Ki respectively *(26,27)*.

In the majority of these studies, SUV_{mean} values have been calculated using manually drawn ROIs or a single fixed threshold (varying from 40% to 75% of the maximum activity). Among these studies only one has considered the reproducibility of metabolic functional volumes, using fixed threshold. Krak et al *(3)* have shown a mean percentage difference in the ROI volumes of 23±20% and 55±35% for a fixed threshold of 50% and 75% respectively. Finally, according to our knowledge there has been no published study evaluating the reproducibility of the TGV/TPV.

To date, despite numerous studies assessing the accuracy of different segmentation algorithms there is a lack of evaluation of the repeatability and reproducibility of these algorithms relative to different threshold and automatic based delineation approaches. Therefore the main objective of this study was to assess the repeatability and reproducibility in determining 3D functional volumes and associated indices (SUV_{mean}, TGV/TPV) in PET imaging using different algorithms. The reproducibility on SUV_{max} was also included since it represents the index mostly used today in clinical practice as well as in order to facilitate a direct comparison with previous studies. This evaluation was carried out on double baseline FDG and FLT clinical PET datasets.

2. Materials and methods

2.1 Segmentation algorithms considered

Four approaches were used in this work. Two different fixed thresholds *(12)* were considered, at 42% (T42) and 50% (T50) of the maximum voxel value, using a region growing algorithm with the maximum intensity voxel as seed.

An adaptive thresholding (TSBR) *(13)* was also included:

$$I_{threshold} = a + b\frac{1}{SBR} \quad (1)$$

SBR is the source-to-background ratio, defined as the contrast between a manually defined background region of interest (ROI) and the mean of the maximum intensity voxel and its eight surrounding neighbours in the same slice. The parameters (a,b) are optimised through linear regression analysis for a given scanner using phantom acquisitions of various sphere sizes and contrast.

For automatic segmentation approaches, the Fuzzy C-Means (FCM) *(28)* clustering algorithm was considered with two clusters (background and lesion). This algorithm has been previously used for functional volume segmentation tasks in both brain and oncology applications *(29,30)* and iteratively minimizes a cost function of the voxels intensity values in order to estimate the centre of each cluster and membership of each voxel to these clusters. The second automatic algorithm considered was the fuzzy locally adaptive Bayesian (FLAB) *(19)* methodology, based on a combination of statistical models with a fuzzy measure in order to simultaneously address both issues of noise and blur resulting from partial volume effects (PVE) in PET images. FLAB is also able to deal with strongly heterogeneous uptake in complex-shaped tumours and generate non binary segmented volumes by considering three classes and the associated fuzzy transitions *(31)*. Estimation of the

parameters required for the segmentation (Gaussian mean and variance of each class and spatial priors for each voxels) are estimated using the iterative Stochastic Expectation Maximization (SEM) procedure. For all approaches, the tumours were delineated after having been isolated in a 3D box of interest previously defined and fixed for all segmentation methodologies (manual and automatic).

2.2 Repeatability, reproducibility: definitions

Within the context of this study repeatability is defined as the ability of a given segmentation algorithm to reach the same result regarding the definition of a functional volume when applied multiple times on a single image. In such a task entirely deterministic fixed threshold approaches (T42,T50) will always give the same result. On the other hand, more advanced methods, like adaptive thresholding or automatic algorithms, such as FCM and FLAB considered here, are susceptible to give different results when applied multiple times on the same image. The adaptive thresholding segmentation, for instance, depends on a manually drawn background ROI and may thus result in variable delineation depending on the choice of this ROI. On the other hand, FCM and FLAB are iterative procedures that may not converge to the same result at each execution. Finally, manual delineation may be considered as the least repeatable, even when considering a single operator (intra-operator variability). A second aspect considered in this study was in terms of the impact of a segmentation algorithm on the reproducibility of determining functional volumes from two baseline PET scans.

Two different clinical datasets were used comprising of esophageal and breast cancer patients scanned with 18F-FDG and 18F-FLT respectively. In both cases, two consecutive PET scans were acquired at few days interval (see section 2.3). We

therefore studied the differences in derived functional tumour volumes, lesion SUV_{mean}, and total glycolytic/proliferative volumes extracted from both images. Repeatability of measuring tumour volumes using the various delineation approaches considered in this study was investigated on the same clinical datasets.

2.3 Validation studies

Fourteen whole body 18F-FDG PET/CT images acquired on patients with esophageal cancer (total of 17 lesions), and nine 18F-FLT PET/CT acquisitions of breast cancer patients (12 lesions total) were considered. Esophageal cancer patients' images were acquired at 3.4±2.2 days interval on a Philips GEMINI PET/CT scanner with 2min acquisition per bed position, 60min after 18F-FDG injection of 6MBq/kg. Data was reconstructed using RAMLA 3D with standard clinical protocol parameters (2 iterations, relaxation parameter of 0.05, 5mm FWHM 3D Gaussian post-filtering). 18F-FLT-PET images were acquired on patients with breast cancer *(27)*, for which two scans were performed within 2-7 days (median 4.1) of each other prior to treatment. All patients received a single bolus intravenous injection of 18F-FLT (153-381 MBq) over 30s, and dynamic PET scanning was performed for 95 min. Patients were scanned on a CTI/Siemens ECAT962/HR+ PET scanner and data was reconstructed using OSEM (360 iterations, 6 subsets, no post-filtering).

In both cases two baseline scans were acquired within an average of 3-4 days from each other. As no treatment was administered between the two baseline scans, and considering the short time between the two acquisitions the assumption is that no significant physiological changes occurred in between. A similar assumption has been previously used in all other studies evaluating the reproducibility and repeatability of different SUV measures in PET imaging with double baseline scans

carried out within 5-10 days from each other *(21-25)*. Figure 1 shows the two baseline scans for one (a) esophageal and (b) breast cancer patient.

2.4 Analysis

For the repeatability evaluation, the tumours in the first image for each patient were segmented ten times each with FCM, FLAB, and TSBR. In addition, manual delineation was carried out by two nuclear medicine experts. More specifically the two experts performed ten different slice-by-slice manual delineations for the different lesions considered in a randomised fashion, ensuring a minimum of a week between two consecutive delineations of the same lesion. All these manual segmentations were carried out under the same conditions of full range contrast display. The mean percentage variability and associated standard deviation with respect to the mean segmented volume was computed for each of the lesions and segmentation approaches across the ten executions and across the ten manual delineations, in order to assess the repeatability of the approaches. The repeatability of the manual delineations from the two experts were compared separately (intra-observer variability) and to each other (inter-observer variability) using intra-class coefficients (ICCs).

To study the relative impact of the different segmentation algorithms on the reproducibility of deriving different PET image indices, tumour volumes were segmented independently on both baseline scan images for each lesion, using all the different automatic segmentation approaches considered (see section 2.1). Subsequently, TV (in cm^3), SUV_{mean}, TGV and SUV_{max} quantitative values M were computed for each delineated lesion and compared between the two scans using the mean percentage difference relative to the mean of both baseline scans:

$$(M_{scan2} - M_{scan1}) / \frac{(M_{scan1} + M_{scan2})}{2} \times 100 \qquad (2)$$

The distribution of the differences between each pair of measurements was assessed for each of the considered index using the Kolmogorov-Smirnov test, showing no significant differences from a normal distribution (see figure 2). Bland-Altman analysis *(32)* was subsequently used to highlight differences between segmentation methodologies. Mean and standard deviation (SD) of differences as well as the respective 95% confidence intervals (CI) were obtained. In order to define the reproducibility limits (normal range of spontaneous changes) the 95% CI for the difference between two measurements were computed as the mean difference ±1.96 times the SD of the difference. In order to investigate any potential correlations in the measured reproducibility the magnitude of the percentage difference for the TV, SUV_{max} and SUV_{mean} measurements were compared to the average of the tumour volumes using Pearson correlation coefficient *r*. This analysis was repeated to investigate the correlation of the reproducibility of the different parameters with the SUV_{mean}.

3. Results

Table I contains the mean variability and standard deviation around the mean segmented volume across the ten manual delineations performed from each of the two nuclear medicine experts, and 10 repeated executions of the FLAB, FCM and TSBR algorithms. Results on both clinical datasets are presented separately. FLAB demonstrated highly repeatable results in all of the studied cases, with negligible variability (1%) around the mean segmented 3D volumes across the different repeated executions. FCM also lead to satisfactory repeatability results (1.4±1.6% for

the FDG cases and 2.3±1.9% for the FLT cases). In comparison, the use of the TSBR led to more than twice as high variability (2.9±2.7% and 4.7±3.6% for the FDG and FLT cases respectively). By contrast manual segmentation by the two experts showed high intra-observer variability for FDG esophageal lesions (14.1±12.1% and 16.4±11.3% for expert 1 and 2 respectively). Inter-observer variability was 17.1±14.3% with an ICC of 0.67 (CI: 0.39-0.89). In the case of FLT, this variability was even higher, with intra-observer variability of 22.1±18.7% and 23.8±17.8% for expert 1 and 2 respectively and an inter-observer variability of 27.4±21.9% with ICC 0.59 (CI: 0.31-0.84).

Tables II and III contain a summary of the reproducibility results for the different parameters computed from Bland-Altman plots on the two consecutive baseline scans for FDG esophageal and FLT breast lesions respectively. The observed reproducibility of SUV_{max} and SUV_{mean} measurements for the volumes obtained using TSBR and FLAB is illustrated in figure 3. The corresponding plots for TV are shown in figures 4(A) and 4(B) using TSBR and FLAB respectively.

Concerning the reproducibility of SUV_{max} similar percentage differences were measured for the FDG and FLT datasets with a SD of the mean percentage difference of 16.7% and 14.9% respectively. The upper and lower percentage reproducibility limits for the SUV_{max} was -31% to 35% and -30% to 28% for the FDG and FLT datasets respectively. On the other hand the automatic approaches led to FDG TV measurement reproducibility limits of -21% to 31% and -51% to 52% for the FLAB and the FCM algorithms respectively. A poorer reproducibility of the FDG TV measurements was observed for the threshold based approaches with upper and lower reproducibility limits of -90% to 51% and -69% to 73% for the adaptive and T42 respectively. In the case of FLT TV measurements, the reproducibility was similar to

FDG for the threshold based approaches, while a deterioration in the reproducibility obtained with the automatic approaches was observed particularly for the FCM algorithm with reproducibility limits of -66% to 74%.

SUV_{mean} measurements using FLAB exhibited reproducibility levels of similar magnitude to that for the TV definition, with a SD of the mean percentage difference of 15.6% and 14.1% for the FDG and FLT datasets respectively. This was however not the case for the other tumour delineation algorithms considered, with the larger SUV_{mean} reproducibility limits using the FCM tumour definition (-77% to 62% and -59% to 59% for the FDG and FLT datasets respectively). Finally, the smaller SUV_{mean} reproducibility for the threshold based approaches was obtained using T50, for both the FDG and FLT datasets with a mean percentage difference of -10.5±23% and -13.3±16.8% respectively.

The reproducibility of TGV/TPV, being the product of TV and SUV_{mean}, was dependent on the direction of changes for both TV and SUV_{mean}. As an increase (respectively decrease) of TV was correlated with a decrease (respectively increase) of SUV_{mean} ($p<0.002$, $r=0.54$, 0.67, 0.72 for FLAB, TSBR and T42 respectively), TGV/TPV reproducibility levels were generally similar in magnitude to the TV and SUV_{mean} considered separately. However, in certain cases there were more increases or decreases of both TV and SUV_{mean} for a given patient, resulting in larger variability of the TGV/TPV measurements (for example the TSBR measurements of the FLT breast lesions, with 22.1±48.9% for the TPV whereas TV and SUV_{mean} were 11.3±31.4% and -3.2±26.5% respectively).

The TV reproducibility results were dependent on the measured TV with a larger variability seen for smaller tumours. This dependence was statistically significant for the adaptive thresholding ($r=0.37$, $p=0.046$, see figure 5(A)) with

differences higher than 30% on average (up to 75%) in several of the tumours below 50cm^3. On the other hand this dependence was not significant for FLAB (r=0.27, p=0.16, figure 5(B)) with most differences <30% irrespective of TV, further demonstrating improved robustness as previously shown *(19,20)*. In terms of the SUV_{max} reproducibility results there was no statistically significant trend with either the lesion size (r=0.016, p=0.93, figure 5(C)) or the mean of the two measured SUV_{mean} values (r=0.14, p=0.49). Finally no statistically significant trends were found for the SUV_{mean} reproducibility depending on the lesion size irrespective of the segmentation algorithm used (r=0.2, p=0.3 and r=0.23, p=0.23 for TSBR and FLAB respectively).

4. Discussion

Functional volume delineation represents today an area of interest for multiple clinical (routine and research) applications of PET imaging (prognosis, response prediction, therapy assessment, radiotherapy treatment planning). In all of these applications, the repeatability and reproducibility with which functional volumes can be determined under different imaging conditions plays a predominant role, allowing a level of confidence to be established in the use of such tumour volume measurements. Volume definition methodologies currently used in clinical practice are based on the use of manual delineation or fixed and adaptive thresholding *(12-14)*, while several promising automatic algorithms have been recently proposed *(16-19)*. The major drawback of manual delineation is high inter- and intra-observer variability in addition to being time consuming. On the other hand, currently considered state of the art adaptive threshold based algorithms have been shown to accurately define functional volumes under certain imaging conditions of spherical

and homogeneous activity distribution lesions. However, adaptive thresholding approaches usually involve some user interaction to select background regions of interest, which can potentially lead to user introduced variability. Although signal intensity reproducibility, predominantly considering the use of SUV_{max}, has been previously assessed, the potential of new indices such as tumour volume and/or TGV/TPV can be only considered following the assessment of their reproducibility which has not been previously widely assessed. Therefore in this study the reproducibility limits of these indices, in comparison to other indices considered as the current gold standard, have been assessed using different tumour delineation methodologies on double baseline FDG and FLT datasets.

In terms of repeatability, all algorithms considered exhibited mean differences <5%, with automatic approaches coming closer to the perfect repeatability that can be achieved by deterministic approaches such as a fixed threshold. The repeatability of both threshold and automatic segmentation approaches was superior to that of manual delineation. This of course should be considered within the context of the limited absolute accuracy of thresholding, particularly for non-homogeneous in form and activity distribution lesions *(31)*.

The variability in the SUV_{max} observed in this work is similar to that measured in previous reproducibility studies, with similar percentage differences for FDG and FLT datasets, suggesting that differences larger than -30% can be considered as significant in treatment response, while changes above 35% (30% for FLT) may be indicative of no response. Depending on the delineation algorithm used, the mean percentage difference and corresponding SD for TV measured on the two baseline scans varied from 5%±13% (4%±16%) to -19%±36% (10%±35%) for the FDG (FLT) datasets. The smallest TV reproducibility limits obtained were similar to those for

SUV_{max}, ranging from -21% to 31% and -27% to 35% for FDG and FLT respectively, suggesting in turn that, depending on the segmentation algorithm used, similar to SUV_{max} confidence intervals may be considered for monitoring therapy response based on functional TV. Similarly in the case of TGV/TPV the smallest reproducibility limits measured were between -16% to 26% and -30% to 37% for FDG and FLT respectively. On the other hand, the largest reproducibility limits for the FDG TV and TGV ranged from -90% to 73% and -68% to 52% respectively.

Reproducibility ranges obtained on the FDG esophageal lesions were almost systematically smaller than the ones obtained on the FLT breast lesions dataset, which can be attributed to the higher level of noise and overall lower contrast observed in the FLT cases, resulting in less robust delineations. In addition, FDG esophageal lesions tended to appear more homogeneous than breast lesions. For instance, FCM which incorporates neither noise nor spatial modelling is associated with a larger mean TV variability on the FLT dataset relative to FDG, whereas FLAB exhibits similar reproducibility levels for both. This highlights the need for a robust delineation tool ensuring high reproducibility in an environment of substantial image quality variability, likely for example to be encountered in multi-center trials where the use of functional TV as a measure of response to therapy may be considered.

T50 uses a more restrictive threshold than 42% and is therefore less prone to large over-evaluation of low contrast (<4:1) and/or small size (<2cm in diameter) tumor volumes. It led to systematically lower variability than T42. Finally, the adaptive thresholding methodology did not demonstrate better reproducibility than fixed thresholding, which can be attributed to the use of the background ROI placed manually on both scans, combined with the fact that background activity may also vary between the two scans.

Although a potential criticism for the current study can be the lack of ground-truth for the functional volumes, the aim of this work was not to assess the absolute accuracy of algorithms, which has been previously assessed for the approaches used in this work *(19,31)*. The objective was to assess the reproducibility limits of functional volume related indices that can be attained depending on the algorithm. Within this context, the repeated studies of the double baseline acquisitions have been performed within an average of 3-4 days from each other without any treatment between them, matching that used by all other reproducibility studies to date *(21-25)*. Finally the reproducibility of the SUV_{max} was included in this work as the current gold standard facilitating at the same time the comparison of our reproducibility study to those performed previously. The SUV_{max} reproducibility limits obtained in this work for both FDG and FLT agree closely with those of previous studies.

5. Conclusion

The smaller reproducibility ranges obtained for the different image indices considered in this study, similar to those of SUV_{max}, suggest that new automatic segmentation approaches may facilitate the introduction of tumour volumes or a combination of tumour volumes and signal intensity in the form of total glycolytic/proliferative volumes derived from PET images for therapy response studies. However, our results also demonstrate that the reproducibility of different quantitative parameters associated with functional volumes depends significantly on the delineation approach.

Acknowledgments

We gratefully acknowledge funding by the Ligue Contre le Cancer (Finistére Committee), the French National Research Agency (ANR-08-ETEC-005-01), Cancéropôle Grand Ouest (R05014NG), and the CR-UK&EPSRC Cancer Imaging Centre (Imperial College, London), the UK Medical Research Council and the Department of Health (C2536/A10337, U.1200.02.005.00001.01).

References

1. Kelloff GJ, Hoffman JM, Johnson B, et al. Progress and promise of FDG PET imaging for cancer patient management and oncologic drug development. *Clin Cancer Res.* 2005;11:2785-2808.

2. Visvikis D, Cheze-Le Rest C, Costa DC, Bomanji J, Gacinovic S, Ell PJ. Influence of OSEM and Segmented Attenuation Correction in the Calculation of Standardised Uptake Values for 18FDG-PET. *Eur J Nucl Med Mol Im.* 2001;28:1326-1335.

3. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Im*. 2005;32:294-301.

4. Lucignani G, Larson S. Doctor, what does my future hold? The prognostic values of FDG PET in solid tumours. *Eur J Nucl Med Mol Im.* 2010, in press.

5. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors. *J Nucl Med*. 2009;50:122S-150S.

6. Seol YM, Kwon BR, Song MK, et al. Measurement of tumor volume by PET to evaluate prognosis in patients with head and neck cancer treated by chemo-radiation therapy. *Acta Oncol*. 2010;49(2):201-8.

7. Chung MK, Jeong HS, Park SG, et al. Metabolic tumor volume of (18F)-fluorodeoxyglucose positron emission tomography/computed tomography predicts short-term outcome to radiotherapy with or without chemotherapy in pharyngeal cancer. *Clin Cancer Res*. 2009;15(18):5861-8.

8. Hyun SH, Choi JY, Shim YM, et al. Prognostic value of metabolic tumor volume measured by 18F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. *Ann Surg Oncol.* 2010;17(1):115-22.

9. Larson SM, Erdi Y, Akhurst T, et al. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET FDG imaging: the visual response score and the change in total lesion glycolysis, *Clin Pos Imag*. 1999;2:159-171.

10. Francis RJ, Byrne MJ, Van der Schaaf AA, et al. Early Prediction of Response to Chemotherapy and Survival in Malignant Pleural Mesothelioma Using a Novel Semiautomated 3-Dimensional Volume-Based Analysis of Serial 18F-FDG PET Scans. *J Nucl Med*. 2007;48:1449-1458.

11. Cazaentre T, Morschhauser F, Vermandel M, et al. Pre-therapy 18F-FDG PET quantitative parameters help in predicting the response to radioimmunotherapy in non-Hodgkin lymphoma. *Eur J Nucl Med Mol Im.* 2010, 37:494–504

12. Erdi E, Mawlawi O, Larson SM, et al. Segmentation of Lung Lesion Volume by Adaptative Positron Emission Tomography Image Thresholding. *Cancer.* 1997;80(S12):2505-2509.

13. Daisne J-F, Sibomana M, Bol A, et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radioth. Oncol*. 2003;69:247-250.

14. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of Different Methods for Delineation of 18F-FDG PET-Positive Tissue for Target Volume Definition in Radiotherapy of Patients with Non-Small Cell Lung Cancer. *J Nucl Med*. 2005;46(8):1342-8.

15. Biehl KJ, Kong FM, Dehdashti F, et al. 18F-FDG PET definition of gross tumor volume for radiotherapy of non-small cell lung cancer: is a single standardized uptake value threshold approach appropriate? *J Nucl Med.* 2006;47:1808-1812.

16. El Naqa I, Yang D, Apte A, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. *Med Phys.* 2007;34(12):4738-4749.

17. Montgomery DWG, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys.* 2007;34(2):722-736.

18. Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Med Mol Im*. 2007;34:1427-1438.

19. Hatt M, Turzo A, Roux C, et al. A fuzzy Bayesian locally adaptive segmentation approach for volume determination in PET. *IEEE Trans Med Im*. 2009;28(6):881-893.

20. Hatt M, Bailly P, Turzo A, Roux C, Visvikis D. PET functional volume segmentation: a robustness study. *IEEE Medical Imaging Conference proceedings*. 2008;4335-4339.

21. Minn H, Clavo AC, Grenman R, Wahl RL. In vitro comparison of cell proliferation kinetics and uptake of tritiated fluorodeoxyglucose and L-methionine in squamous-cell carcinoma of the head and neck. *J Nucl Med*. 1995;36:252-258.

22. Weber WA, Ziegler SI, Thodtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG PET. *J Nucl Med*. 1999;40:1771-1777.

23. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors. *J Nucl Med. 2008*;49:1804-1808.

24. Paquet N, Albert A, Foidart J, Hustinx R. Within patient variability of FDG standardised uptake values in normal tissues. *J Nucl Med.* 2004;45:784-788.

25. Velasquez LM, Boellaard R, Kollia G, et al. Repeatibility of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. *J Nucl Med*. 2009;50:1646-1654.

26. De Langen AJ, Klabbers B, Lubberink M, et al. Reproducibility of quantitative 18FLT measurements using positron emission tomography. *Eur J Nucl Med Mol Im.* 2009;36:389-395.

27. Kenny L, Coombes RC, Vigushin DM, et al. Imaging early changes in proliferation at 1 week post chemotherapy: a pilot study in breast cancer patients with FLT positron emission tomography. *Eur J Nucl Med Mol Im.* 2007;34:1339-1347.

28. Dunn JC. A fuzzy relative of the isodata process and its use in detecting compact well-separeted clusters. *J Cybernet*. 1974;31:32-57.

29. Zhu W, Jiang T. Automation segmentation of PET image for brain tumours. *IEEE MIC proceedings.* 2003;4:2627-2629.

30. Belhassen S, Zaidi H. Segmentation of heterogeneous tumors in PET using a novel fuzzy C-means algorithm. *J Nuc Med.* 2009;50(S2):1442.

31. Hatt M, Cheze-le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int J Radiat Oncol Biol. Phys.* 2010;77:301-308.

32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307-310.

Method	17 esophageal lesions		12 breast lesions	
	Mean variability (%)	Standard deviation	Mean variability (%)	Standard deviation
FLAB	0.6	0.3	1.1	0.7
FCM	1.4	1.6	2.3	1.9
Fixed threshold	0	0	0	0
Adaptive threshold	2.9	2.7	4.7	3.6
Manual delineation (expert 1)	14.1	12.2	22.1	18.7
Manual delineation (expert 2)	16.4	11.3	23.8	17.8
Manual delineation (expert 2 w/r to 1)	17.1	14.3	27.4	21.9

Table I

Method / parameter		% Difference (FDG)					
		Mean ± SD	95% CI	Lower reproducibility limit, LRL	95% CI for LRL	Upper reproducibility limit, URL	95% CI for URL
SUV_{max}		1.8 ± 16.7	-6.8 to 10.4	-30.9	-45.9 to -16	34.6	19.9 to 49.6
FLAB	TV	5 ± 13.3	-1.8 to 11.9	-21.1	-33 to -9.1	31.1	19.2 to 43
	SUV_{mean}	0 ± 15.6	-8 to 8	-30.5	-44.4 to -16.6	30.5	16.5 to 44.4
	TGV	5.1 ± 10.6	-0.4 to 10.5	-15.8	-25.3 to -6.3	25.9	16.4 to 35.5
FCM	TV	0.4 ± 26.4	-13.2 to 14	-51.4	-75.1 to -27.7	52.2	28.5 to 75.9
	SUV_{mean}	-7.8 ± 35.5	-26 to 10.5	-77.4	-109.2 to -45.5	61.8	30 to 93.7
	TGV	-7.4 ± 30.2	-22.9 to 8.2	-66.6	-93.7 to -39.5	51.9	24.8 to 78.9
TSBR	TV	-19.4 ± 36	-37.9 to -0.9	-89.9	-122.1 to -57.6	51.1	18.9 to 83.3
	SUV_{mean}	6.3 ± 27.4	-7.8 to 20.4	-47.4	-72 to -22.8	60.1	35.5 to 84.6
	TGV	-13 ± 28.2	-27.5 to 1.5	-68.2	-93.4 to -42.9	42.2	17 to 67.4
T_{42}	TV	2.1 ± 36.1	-16.5 to 20.7	-68.7	-101.2 to -36.3	72.9	40.5 to 105.3
	SUV_{mean}	-10.5 ± 30	-25.9 to 5	-69.3	-96.2 to -42.4	48.4	21.5 to 75.3
	TGV	-8.4 ± 23.4	-20.5 to 3.6	-54.3	-75.3 to -33.3	37.5	16.5 to 58.5
T_{50}	TV	0.9 ± 32.9	-16 to 17.8	-63.5	-92.9 to -34	65.3	35.9 to 94.8
	SUV_{mean}	-10.5 ± 23	-22.6 to 1.6	-56.5	-77.6 to -35.5	35.6	14.5 to 56.6
	TGV	-9.5 ± 23.1	-21.4 to 2.4	-54.9	-75.6 to 34.1	35.8	15.1 to 56.6

Table II

Method / parameter		% Difference (FLT)					
		Mean ± SD	95% CI	Lower reproducibility limit, LRL	95% CI for LRL	Upper reproducibility limit, URL	95% CI for URL
SUV_{max}		-0.9 ± 14.9	-10.4 to 8.5	-30	-46.6 to -13.4	28.2	11.6 to 44.8
FLAB	TV	4.3 ± 15.7	-5.7 to 14.3	-26.5	-44.1 to -8.9	35.2	17.6 to 52.8
	SUV_{mean}	-0.6 ± 14.1	-9.6 to 8.3	-28.2	-44 to -12.5	27	11.2 to 42.7
	TPV	3.7 ± 17.2	-7.2 to 14.6	-30	-49.2 to -10.8	37.4	18.2 to 56.6
FCM	TV	4.2 ± 35.7	-18.4 to 26.9	-65.6	-105.5 to -25.8	74.1	34.3 to 114
	SUV_{mean}	0.3 ± 30.1	-18.8 to 19.4	-58.6	-92.2 to -25	59.2	25.6 to 92.8
	TPV	4.6 ± 29.8	-14.3 to 23.6	-53.9	-87.2 to -20.5	63.1	29.7 to 96.4
TSBR	TV	11.3 ± 31.4	-8.7 to 31.2	-50.4	-85.5 to -15.2	72.8	37.7 to 108
	SUV_{mean}	-3.2 ± 26.5	-20 to 16.6	-55.1	-84.7 to -25.5	48.7	19.1 to 78.3
	TPV	22.1 ± 48.9	-9 to 53.2	-73.8	-128.5 to -19.1	118	63.3 to 172.7
T_{42}	TV	9.8 ± 35	-12.4 to 32.1	-58.7	-97.8 to -19.6	78.4	39.3 to 117.5
	SUV_{mean}	-9.4 ± 20.9	-22.7 to 3.9	-50.3	-73.7 to -27	31.6	8.2 to 54.9
	TPV	0.7 ± 27.3	-16.7 to 18	-52.8	-83.3 to -22.3	54.1	23.6 to 84.6
T_{50}	TV	11.2 ± 31.4	-8.8 to 31.1	-50.5	-85.6 to -15.3	72.8	37.6 to 107.9
	SUV_{mean}	-13 ± 16.8	-24 to -2.7	-46.2	-64.9 to -27.4	19.5	0.8 to 38.3
	TPV	-1.8 ± 26	-18.4 to 14.7	-52.8	-81.9 to -23.7	49.1	20.1 to 78.2

Table III

Table captions

Table I: Repeatability evaluation: mean variability and standard deviation around the mean segmented volume for repeated delineations of 17 esophageal and 12 breast lesions on the first baseline FDG and FLT scans respectively.

Table II: Reproducibility results concerning the FDG esophageal lesions with differences of scan #2 measurements with respect to scan #1.

Table III: Reproducibility results concerning the FLT breast lesions with differences of scan #2 measurements with respect to scan #1.

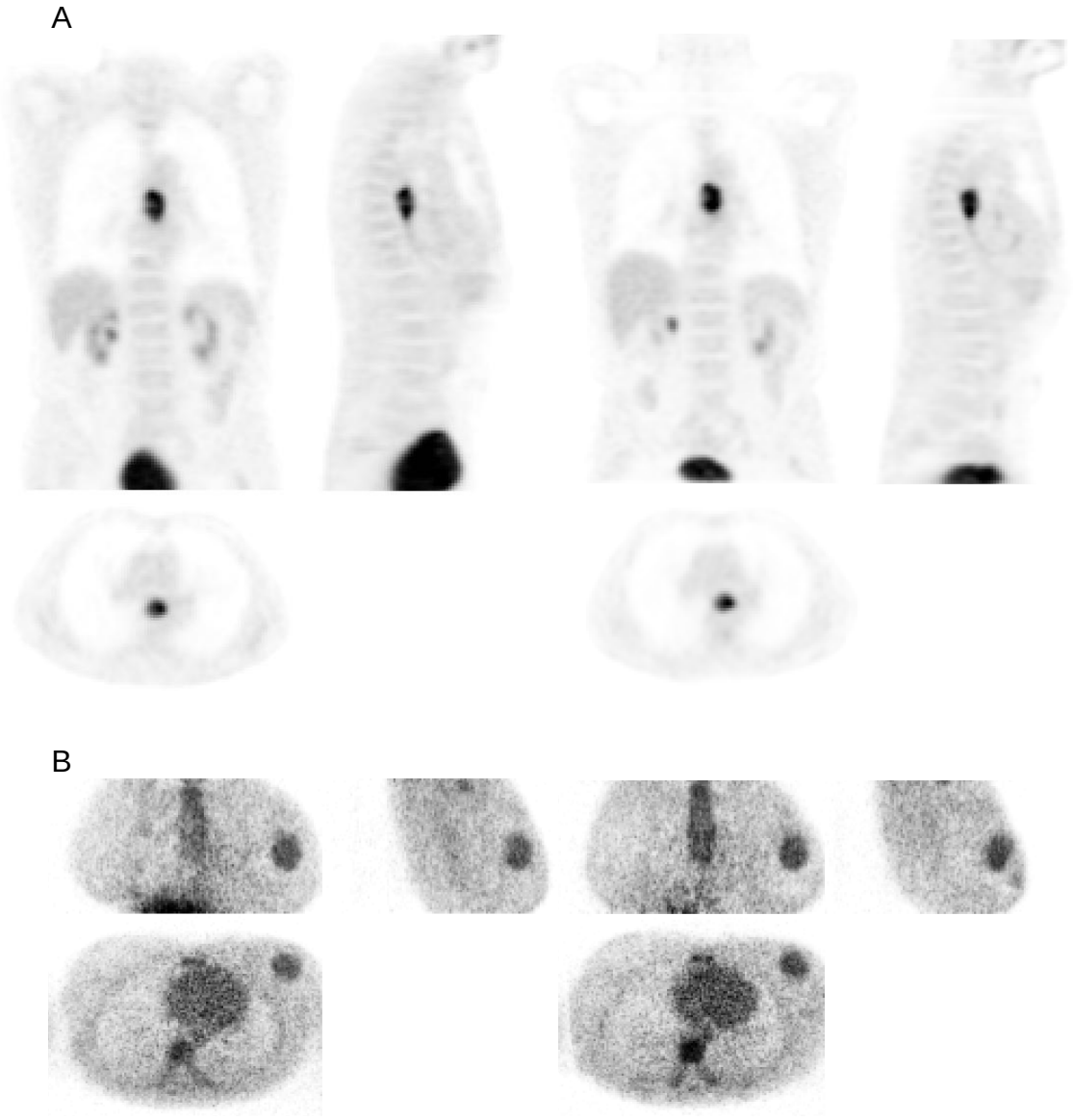


Figure 1

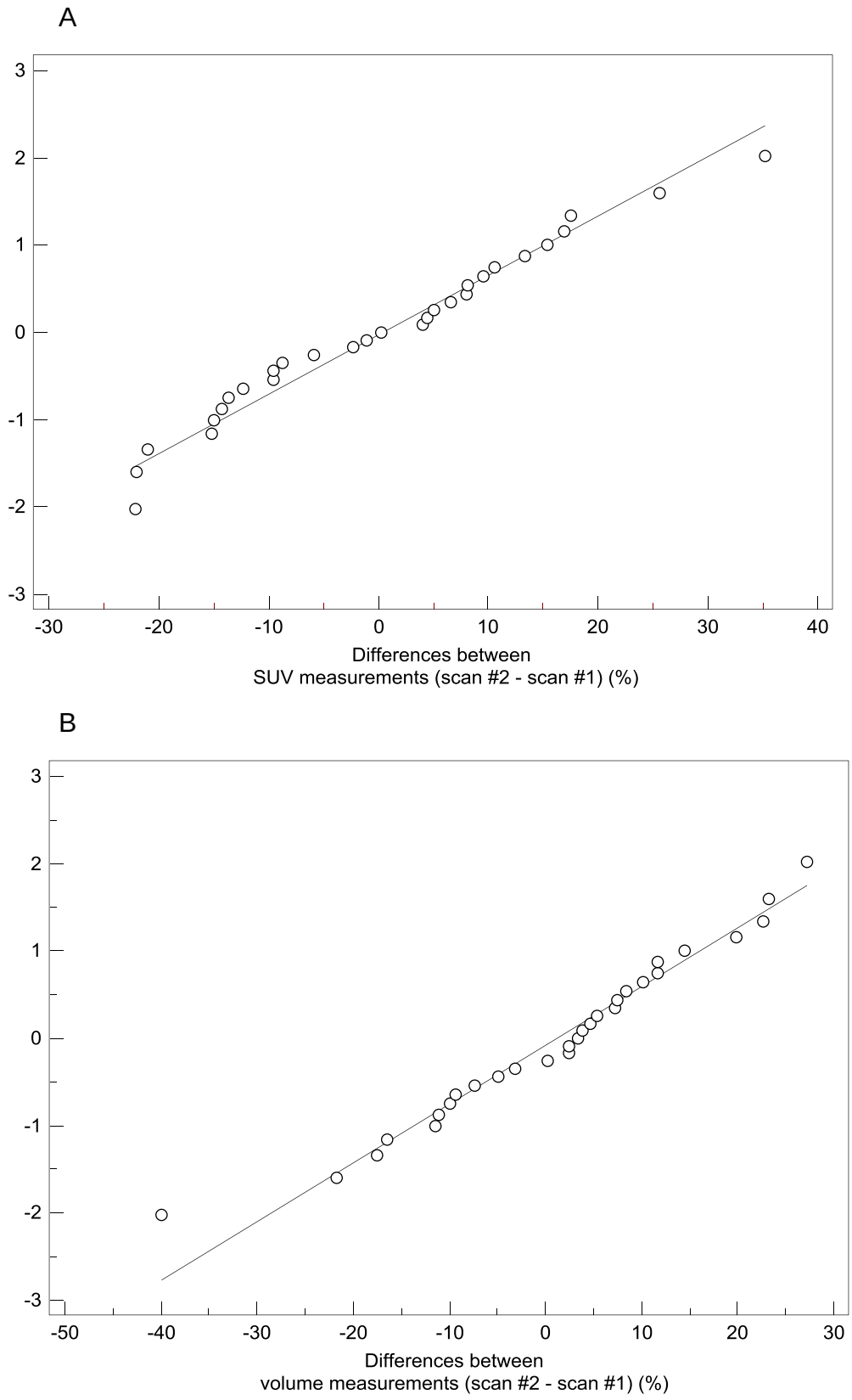
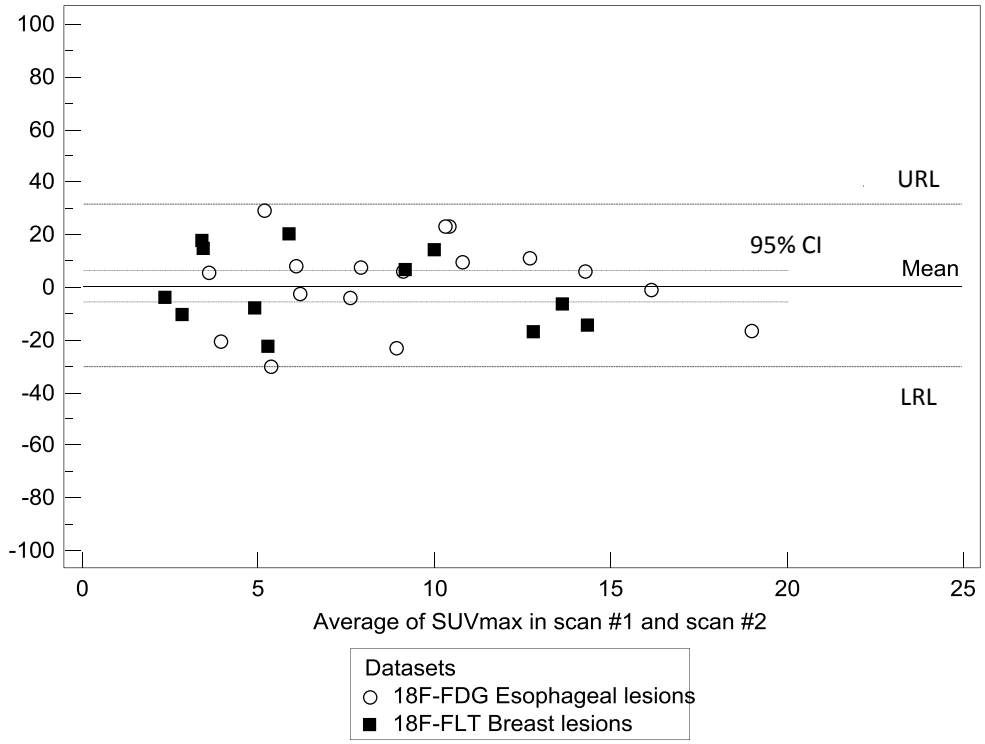
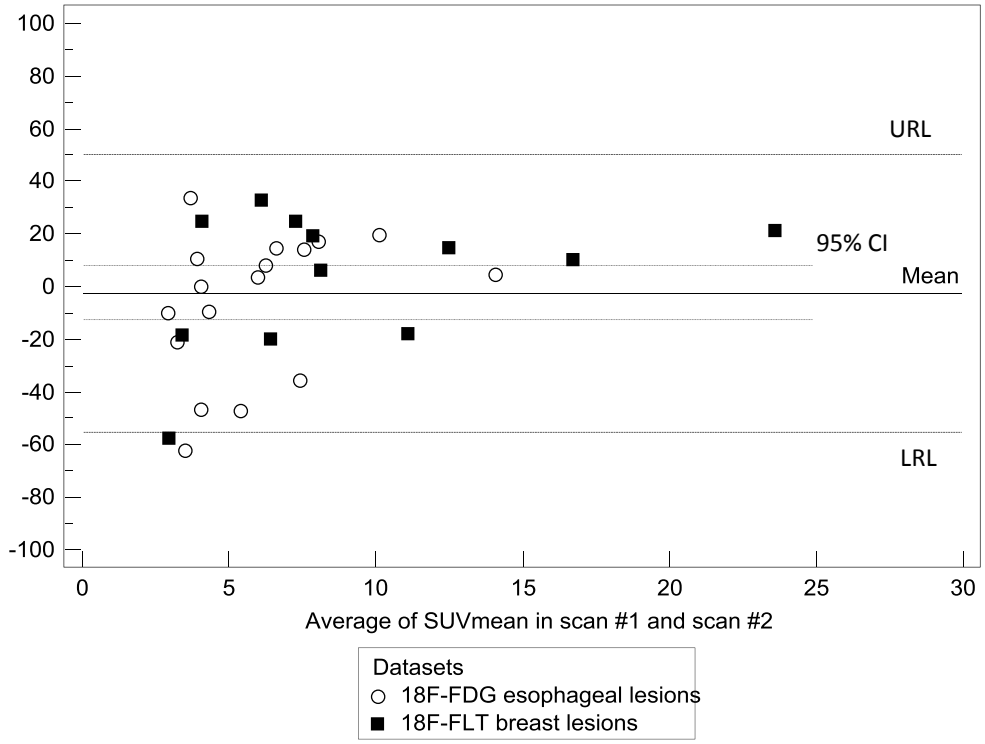

A
3
2
1
0
-1
-2
-3
-30
-20
-10
0
10
20
30
40
Differences between
SUV measurements (scan #2 - scan #1) (%)
B
3
2
1
0
-1
-2
-3
-50
-40
-30
-20
-10
0
10
20
30
Differences between
volume measurements (scan #2 - scan #1) (%)

Figure 2

A


100
80
60
40
20
0
-20
-40
-60
-80
-100
URL
95% CI
Mean
LRL
0
5
10
15
20
25
Average of SUVmax in scan #1 and scan #2
Datasets
18F-FDG Esophageal lesions
18F-FLT Breast lesions

B


100
80
60
40
20
0
-20
-40
-60
-80
-100
URL
95% CI
Mean
LRL
0
5
10
15
20
25
30
Average of SUVmean in scan #1 and scan #2
Datasets
18F-FDG esophageal lesions
18F-FLT breast lesions

C

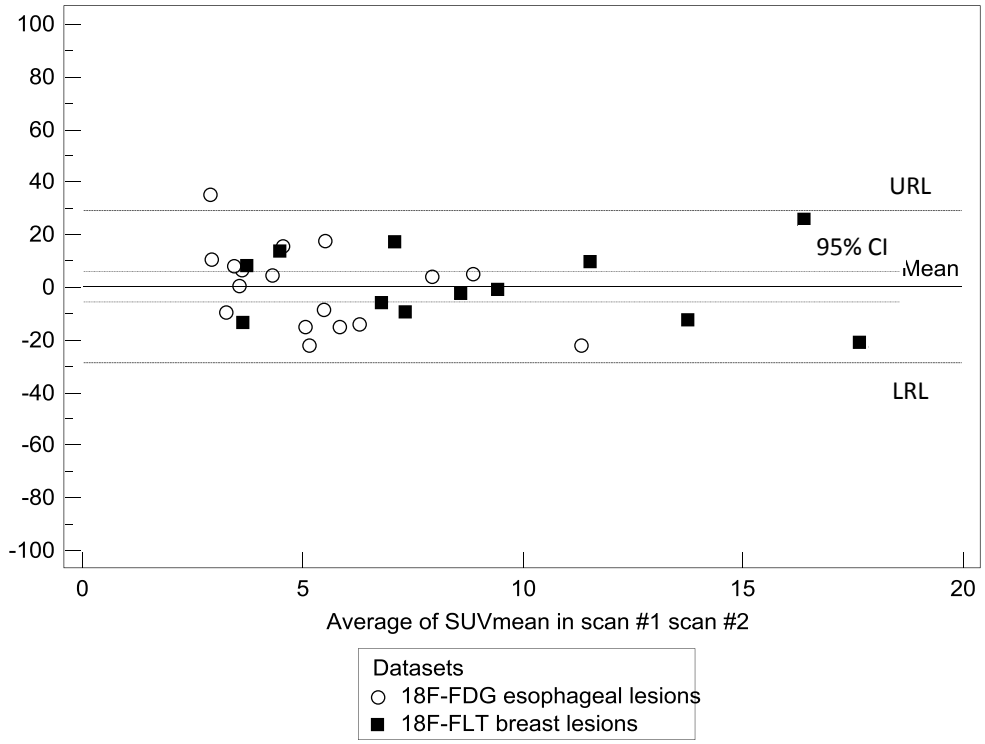
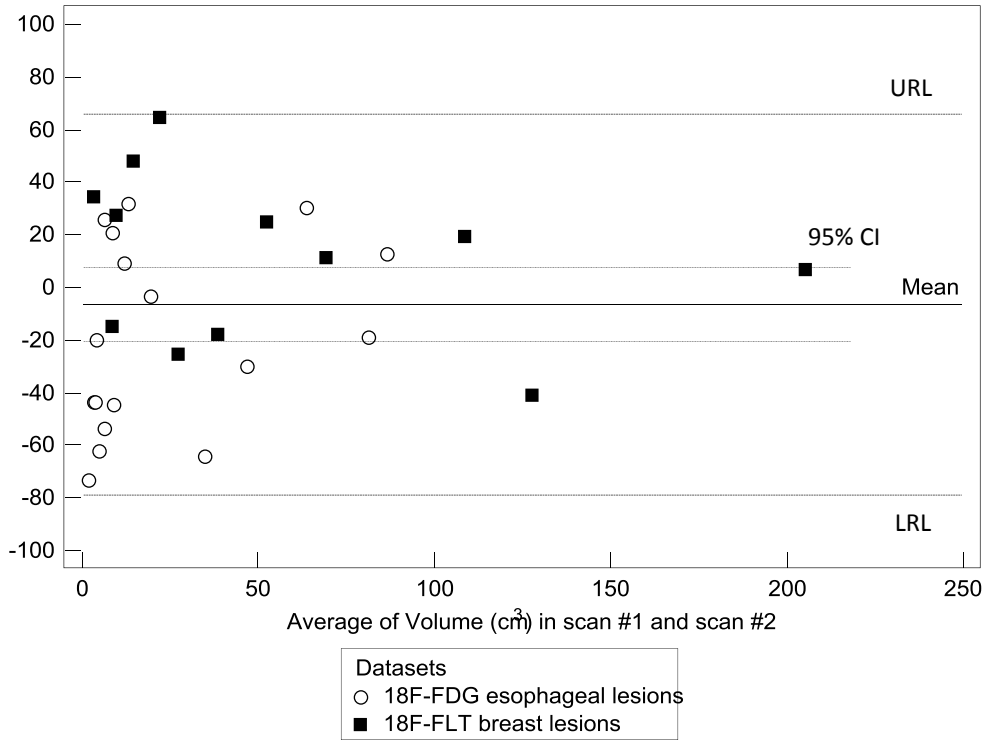

100
80
60
40
20
0
-20
-40
-60
-80
-100
URL
95% CI
Mean
LRL
0
5
10
15
20
Average of SUVmean in scan #1 scan #2
Datasets
18F-FDG esophageal lesions
18F-FLT breast lesions

Figure 3

A


100
80
60
40
20
0
-20
-40
-60
-80
-100
URL
95% CI
Mean
LRL
0
50
100
150
200
250
Average of Volume (cm3) in scan #1 and scan #2
Datasets
18F-FDG esophageal lesions
18F-FLT breast lesions

B

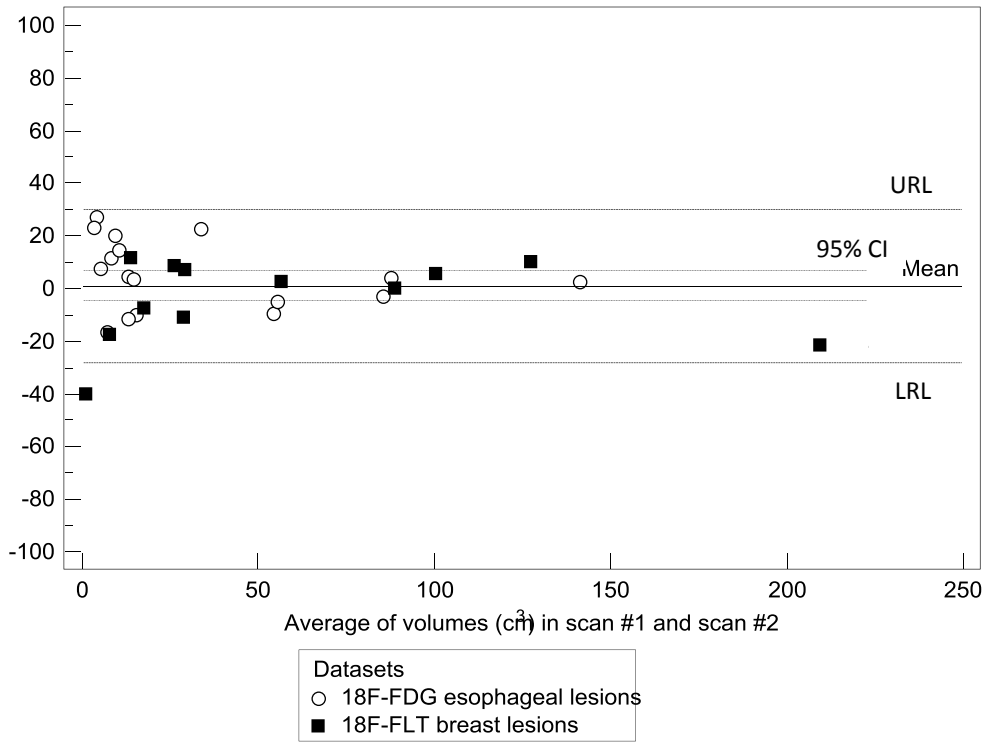
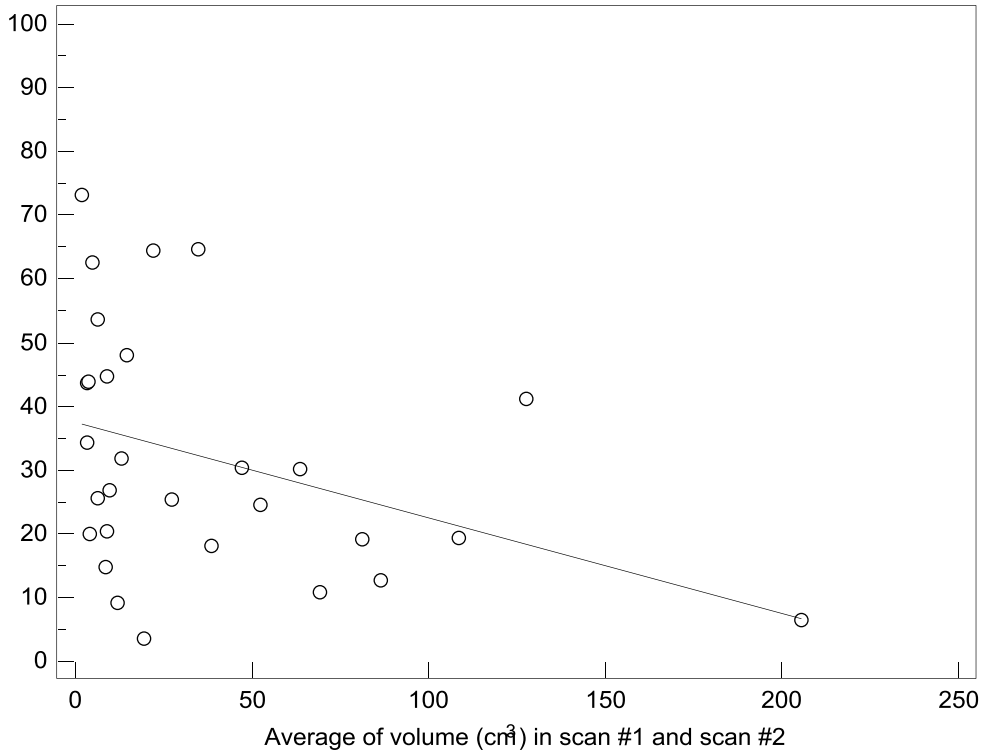
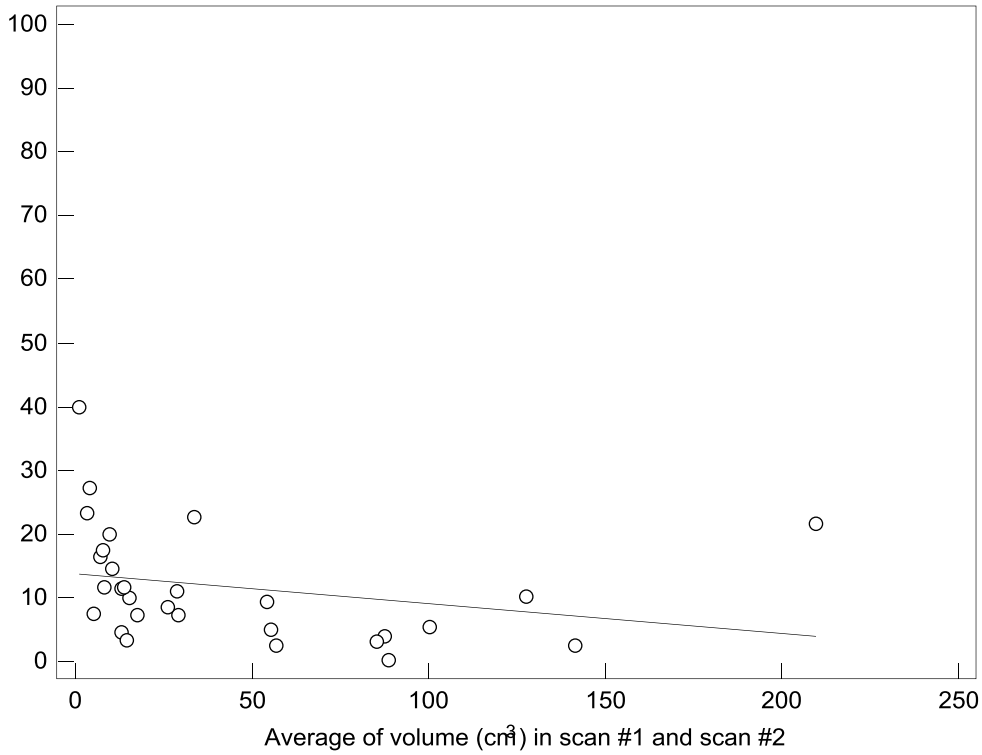

100
80
60
40
20
0
-20
-40
-60
-80
-100
URL
95% CI
Mean
LRL
0
50
100
150
200
250
Average of volumes (cm3) in scan #1 and scan #2
Datasets
18F-FDG esophageal lesions
18F-FLT breast lesions

Figure 4

A


100
90
80
70
60
50
40
30
20
10
0
0
50
100
150
200
250
Average of volume (cm3) in scan #1 and scan #2

B


100
90
80
70
60
50
40
30
20
10
0
0
50
100
150
200
250
Average of volume (cm3) in scan #1 and scan #2

C

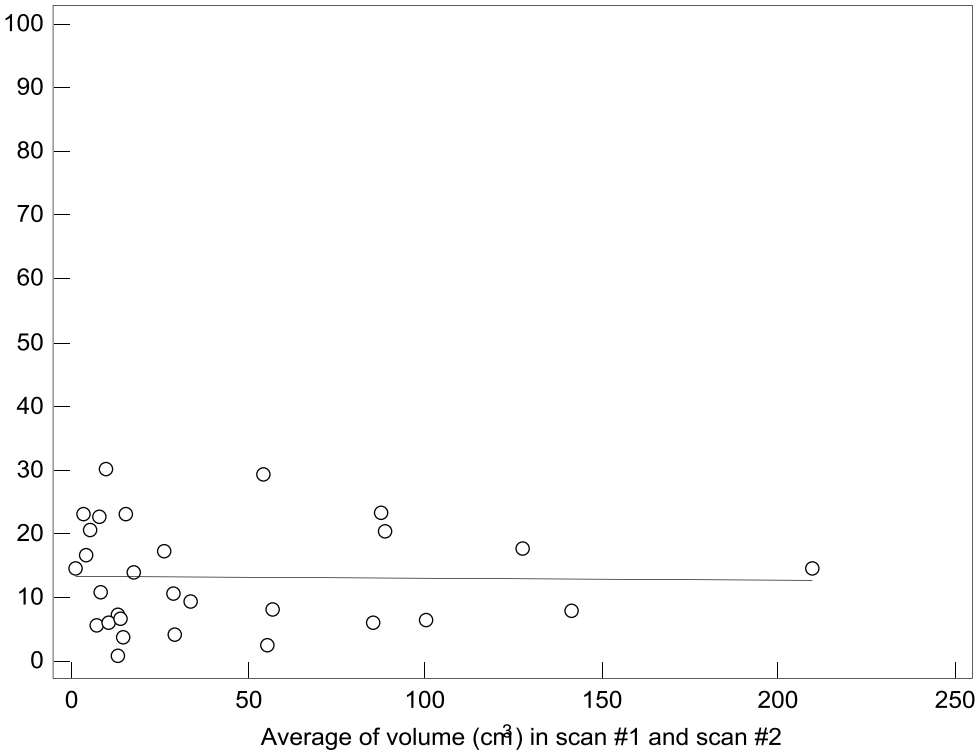

100
90
80
70
60
50
40
30
20
10
0
0
50
100
150
200
250
Average of volume (cm3) in scan #1 and scan #2

Figure 5

Figure Captions

Figure 1: Example of two baseline images (A). FDG (esophagus) and (B). FLT (breast).

Figure 2: Normal plots showing that the distributions of differences for (A). SUV_{mean} (FLAB), and (B). TV (FLAB) between two scans are not significantly different from normality.

Figure 3: Bland-Altman plots of (A). SUV_{max}, (B). SUV_{mean} (adaptive thresholding), and (C). SUV_{mean} (FLAB) values for both FDG and FLT lesions. The lines show the combined mean, 95% CI as well as upper and lower reproducibility limits. Individual values for the FDG and FLT lesions are shown in tables II and III respectively.

Figure 4: Bland-Altman plots of (A). TV (adaptive thresholding), and (B). TV (FLAB) for both FDG and FLT lesions. The lines show the combined mean, 95% CI as well as upper and lower reproducibility limits. Individual values for the FDG and FLT lesions are shown in tables II and III respectively.

Figure 5: Differences between (A-B). tumor volumes and (C). SUV_{max} measured in two baseline scans in relation to the average tumour volume obtained using adaptive thresholding (A), and FLAB (B-C).